Distance, Similarity and Sequence Comparison

Elzinga, C.H.

published in
Advances in Sequence Analysis: Theory, Method, Applications
2014

DOI (link to publisher)
10.1007/978-3-319-04969-4_4

Chapter 4
Distance, Similarity and Sequence Comparison

Cees H. Elzinga

Introduction

Over the last decades, sequence analysis has developed from a fiercely debated trick from computational biology—and I was one of the discussants too—into a broadly accepted toolbox for those interested in classifying all sorts of career data. Recently, with the introduction of ANOVA-like techniques to explain distances in terms of one or more categorical covariates (Bonetti et al. 2013; Studer et al. 2011), sequence analysis tools can be used to test hypotheses about causal relations.

This very book is crammed with state-of-the-art applications of sequence analysis and all chapters have in common that they start by somehow constructing distances and/or similarities from sequence data. Therefore, it seems justified that this paper does not deal with data, not even with simulated data, but instead revisits the fundamental concepts used—distance and similarity—in order to stress the importance of the axiomatic foundations of these concepts. From these axiomatic foundations, we will try to clarify some principles and common misunderstandings pertaining to transforming and normalizing our numerical basis and make some remarks on the relation between the concepts of similarity and distance. None of the ideas in this chapter is new, nor are their interrelations. The purpose of this paper is not to develop new maths or methodology. Instead, its purpose is to make the mathematical concepts accessible and intelligible to social scientists. Therefore, the tone is informal, definitions are sometimes replaced by small graphs and proofs are omitted. On the other hand, since the subject matter is formal and abstract, it is inevitable that I use some formulas and inequalities: I tried to restrict myself to a level of abstractness that allows me to formulate some proposals that are directly applicable to problems of sequence comparison.

We start out in the next two sections to discuss the formal basis of distance and similarity. In Sect. 4, we will discuss the transformations that, respecting the axiomatic basis, may be applied to distance and similarity. In Sect. 5, we discuss the

C. H. Elzinga (✉)
VU University Amsterdam, Amsterdam, The Netherlands
e-mail: c.h.elzinga@vu.nl

Ph. Blanchard et al. (eds.), *Advances in Sequence Analysis: Theory, Method, Applications*, Life Course Research and Social Policies 2, DOI 10.1007/978-3-319-04969-4_4,

relation between the two kinds of measures and deal with the issue of normalization. In Sect. 6, we concisely discuss the fundamental difference between the two concepts as, in practice, partitioning a set of sequences on the basis of similarity may not yield the same partitioning when one would use a distance measure instead. Finally, we summarize in Sect. 7.

Distance

In this section, we will deal with the concept of "distance" and some of the abstract properties of the various measures that we use to evaluate distance. All of us know distances as numbers that refer to the relative location of objects in some space. In this chapter, we do not deal with objects of our everyday lives, located in the physical space that surrounds us, but with quite different objects—sequences—that have no physical location. Hence, we need a definition of "space" that is quite general and that allows us to evaluate distances between sequences according to the same principles that we use in considering distances in our everyday lives.

Axioms of Distance

We shall say that a space consists of a set of objects that has some structure, i.e. a set of rules that govern the relations between the objects. Such relations could be, for example, relations of order or adjacency. In the present context, our space will consist of a set of sequences and the structure will be determined by some distance measure that is defined on all pairs of sequences from that set. However, it is not at all clear what a distance between objects is like, not even in our everyday life. This is illustrated by the two panels of Fig. 4.1. In the left panel, we define the distance between locations a and b as the length of the straight line between these locations. Using the coordinate vectors $\mathbf{a} = (a_1, a_2)$ and $\mathbf{b} = (b_1, b_2)$, the Pythagorean formula yields

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} \tag{4.1}$$

In the right panel however, there is an obstacle, say a building, between a and b. For you, bound to walk the streets, the above calculation of distance is not very relevant so you would rather use

$$d(a,b) = |a_1 - b_1| + |a_2 - b_2|, \tag{4.2}$$

calculating the length of the walking route from a to b. Another, frequently used way of calculating distance is according to the rules of spherical geometry in which shortest distances are not straight lines but curves on the surface of a slightly flat-

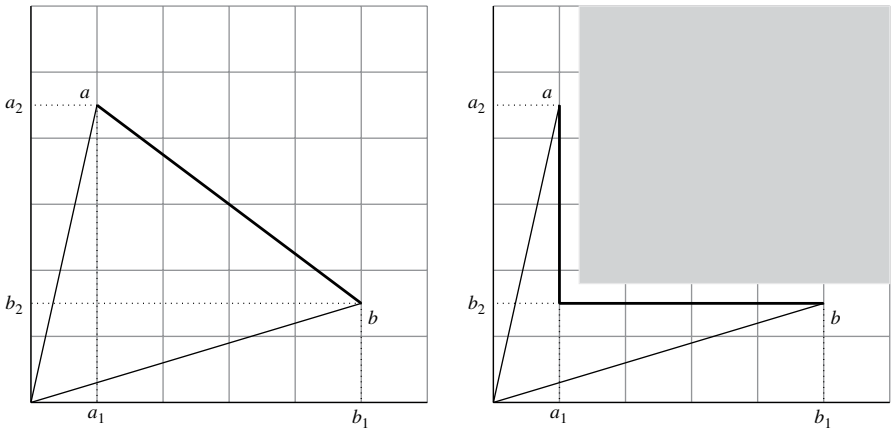

Fig. 4.1 In the left panel, the distance $d(a,b)$ is calculated according to the Pyhagorean formula $d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$ and in the right panel, distance is calculated according to $d(a,b) = |a_1 - b_1| + |a_2 - b_2|$

tened sphere. This is what navigators of ships and planes do when they have to cross long distances, as for example between the ports of Boston and Rotterdam.

So, even in our daily lives, distances are calculated in different ways, depending on their intended use and the properties of the space. Therefore, instead of listing all possible measures of distance, we characterize a function on pairs of elements of a set $X = \{u, v, w, x, \dots\}$ as a distance measure through a few very general properties, also called "axioms". However, these axioms still correspond to our intuitive "living" of distances among cities, buildings and objects in kitchens and offices. Below, we *first* list the axioms and then comment on them. We write $d(u,v)$ for the distance between objects u and v and say that d is a distance if, for all triples u, v, w from the set $X = \{u, v, w, x, \dots\}$, it is true that

$$
\begin{aligned}
&D1: \quad d(u,v) = 0 \text{ if and only if } u = v,\\
&D2: \quad d(u,v) > 0 \text{ if and only if } u \neq v,\\
&D3: \quad d(u,v) = d(v,u),\\
&D4: \quad d(u,w) \leq d(u,v) + d(v,w).
\end{aligned}
$$

A function d that satisfies all four of the above axioms is called a *metric* or, equivalently, a *metric distance* and the pair (X,d) is called a *metric space*. Today, the concept of a metric space is over a hundred years old as it was first coined by Fréchet (1906), at the beginning of the twentieth century, in the context of metricizing the constructive Euclidean geometry.

Axiom D1 says that no two distinct objects can be on the same location and Axiom D2 says that distinct objects must be in different locations. From Axioms D1 and D2, it is immediate that distances cannot be negative. Axiom D3 states that

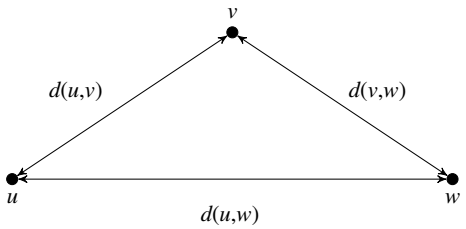

Fig. 4.2 An illustration of the triangle inequality (Axiom D4 in the main text)

distances are symmetric: u is as remote from v as v is remote from u. Clearly, the first three axioms correspond to our intuitions about space and distance. The fourth axiom is called the "triangular inequality", since it pertains to three objects, and it expresses our intuition that "a detour always takes more time". Axiom D4 says that, in order that some measure is to be called a "distance", it should always result in the conclusion that going directly from u to w yields a distance $d(u,w)$ that does not exceed the distance $d(u,v)$ that results from first visiting v plus the distance $d(v,w)$ that results from subsequently traveling from v to w. This is illustrated in Fig. 4.2. A slightly different interpretation is that when two objects (u and w) are close to a third object (v), they must be close to each other, i.e. $d(u,w)$ must be small.

So we see that the triangular inequality fits within our intuitions about space and distance too. Furthermore, given the objects, the triangular inequality acts a a sort of boundary on the numerical values of the metric: $d(u,w)$ is bounded by all pairs $(d(u,x),d(x,w))$, i.e. Axiom D4 must hold for all triples of objects of X. So, Axiom D4 ensures that the metric space is "smooth in all directions" in the sense that if we know the properties of a subspace of a metric, i.e. d on a subset of X, we can be sure that in overlapping sets, these properties will be quite similar. Therefore, Axiom D4 is a very important axiom: if it is not satisfied, it is very risky to generalize from our measurements since we are not sure that the measurements would not be wildly different, had we observed only slightly different objects. So we should avoid the use of "proximities" or "dissimilarities" that are not proper metrics in the sense that they do not satisfy all four axioms D1–D4.

Clearly, the axioms D1–D4 do not prescribe a specific way of measuring distances; the axiom system only limits our freedom of choice of a particular method of gauging distances. In practice, constructing measures that satisfy the first three axioms rarely appears to be a problem; however, constructing measures that also satisfy the triangular inequality is a bit harder. We will come back to this problem in the subsection on normalizing a metric.

OM-Metrics

In all of the chapters of this book, some variant of Optimal Matching (OM) is used to determine distances between sequences. Here we concisely discuss those variants

of OM that generate proper metrics in the sense that the numbers produced satisfy the axioms D1–D4. Formally, let $x = x_1 \dots x_n$ and $y = y_1 \cdots y_n$ denote two n-long state sequences over the alphabet of states $\Sigma = \{\lambda, a, b, \dots\}$ with λ denoting the empty state and let $e = e_1 \dots e_k$ denote a series of admissible sequence edits such that $e(x) = e_k(e_{k-1}(\dots e_2(e_1(x)) \dots)) = y$. For any pair of sequences, there may exist many distinct series of edits that transform x into y and we write $E(x, y)$ to denote the set of such edit-series. Furthermore, to each edit e_i, a nonnegative cost or weight $c(e_i)$ is assigned and the cost of an edit-series $C(e)$ equals the sum of the costs of the edits involved: $C(e) = \sum_i c(e_i)$. The OM-distance $d_{OM}(x, y)$ between a pair of sequences x and y is the minimum of the costs of the edit-series in $E(x, y)$:

$$d_{OM}(x, y) = \min\{C(e) : e \in E(x, y)\}. \tag{4.3}$$

Let us now denote the edit-costs with respect to the characters or states of the alphabet $\Sigma = \{\lambda, a, b, c, \dots\}$ as a symmetric array $\mathbb{C}$ over all pairs of states. In this array, $\mathbb{C}(a, b)$ denotes the cost of substituting a for b, $\mathbb{C}(\lambda, a)$ denotes the cost of deleting a (and substitute it for the empty state λ), and $\mathbb{C}(a, \lambda)$ denotes the cost of inserting state a. With this notation, the "standard" cost matrix is of the form

	λ	a	b	c	…
λ	0	1	1	1	…
a	1	0	2	2	…
b	1	2	0	2	…
c	1	2	2	0	…
⋮	⋮	⋮	⋮	⋮	⋱

A proof that d_{OM} is a proper metric, provided that the cost-matrix itself is a metric can be found in e.g. Yujian and Bo (2007). The condition that the array $\mathbb{C}$ constitutes a metric means that $\mathbb{C}$, understood as a function on pairs of states, satisfies the axioms D1-D4. The reader easily verifies that the axioms indeed hold for the standard cost function as shown above. However, it is not difficult to construct a cost-matrix that violates the triangular inequality D4 as demonstrated in the array below;

	λ	a	b	c
λ	0	1	1	1
a	1	0	1.5	4
b	1	1.5	0	2
c	1	4	2	0

In this array, we have that $4 = \mathbb{C}(a, c) > \mathbb{C}(a, b) + \mathbb{C}(b, c) = 1.5 + 2 = 3.5$, a violation of axiom D4. So, if we do not properly set the edit-costs, the OM-algorithm

will generate numbers that are not proper distances. The condition of a metric cost matrix also implies that OM-variants that use a dynamic, data-driven cost-matrix (Halpin 1950; Hollister 2009; Rohwer and Pötter 1999) will not automatically generate a proper metric over the pertaining sequences.

Rather than directly using the metric properties of a sequence-space, Massoni et al. (2009) tried to utilize the topological structure of an OM-generated sequences space through exploring Kohonen-maps. Perhaps this will turn out to be a seminal approach.

Subsequence-Based Metrics

Once we can represent sequences as vectors in a vector-space, we can use a whole family of proper distance metrics of which the examples in Eqs. 4.1 and 4.2 are just special cases. Let us suppose that we can represent sequences x and y as vectors, i.e. as coordinate arrays[1] $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \dots)$ and $\mathbf{y} = (\mathrm{y}_1, \mathrm{y}_2, \dots)$, it is easy to calculate distances through the Pythagorean formula

$$d(x, y) = \sqrt{\sum_i (\mathrm{x}_i - \mathrm{y}_i)^2} \tag{4.4}$$

$$= \mathbf{x}'\mathbf{x} + \mathbf{y}'\mathbf{y} - 2\mathbf{x}'\mathbf{y}, \tag{4.5}$$

wherein $\mathbf{x}'\mathbf{y} = \sum_i \mathrm{x}_i \mathrm{y}_i$. But how to construct sequence representing vectors? Well, we can do this anyway we like as long as the procedure results in equally long arrays of numbers, one for each sequence. We may select any number of distinct, quantifiable features of the sequences that we consider as relevant, say their length L, their number of distinct states N and the number of spells S, and use the values of these features as the coordinate values of the vectors. In this example we have three features $L(x)$, $N(x)$ and $S(x)$ which can be used to represent each sequence x as a 3-dimensional array

$$\mathbf{x} = (L(x), N(x), S(x)).$$

For example, for the toy-sequence $x = aababbca$, this would yield $\mathbf{x} = (8, 3, 6)$. Obviously, most choices of coordinate-systems will not lead to relevant, useable representations. However, in 2003, I argued (Elzinga 2003, 2005) that using the subsequences as features would generate meaningful results and that idea has been successfully applied in several contexts, e.g. in Berghammer (2010), in Fasang (2010), and in Manzoni et al. (2010). Therefore, we begin with elaborating on the

[1] Please note that I write x, y for sequences, x_1, y_1 for the states of the sequences, $\mathbf{x}$, $\mathbf{y}$ for the representing vectors and $\mathbf{x}_1$, $\mathbf{y}_1$ for the coordinates of the vectors.

concept of "subsequence". For a more formal treatment, the reader is referred to e.g. Crochemore et al. (2007) or Elzinga et al. (2013).

Consider the toy-sequence $x = x_1x_2x_3x_4 = abac$ over the state alphabet $\Sigma = \{\lambda, a, b, c\}$. We may take any nonnegative number of states from x and we will then be left with a subsequence of x: a subsequence u of states that have the same order in x and we will write $u \sqsubseteq x$ to denote such fact. For example, when we take out the a's from x, we will be left with $u = bc$, one of the five 2-long subsequences of x. At most, we can take away all states from x and we will then be left with the empty sequence λ. We might also take the smallest nonnegative number of states from x, zero states, and we would be left with x itself and hence we conclude that $x \sqsubseteq x$. The reader easily verifies that x has 13 distinct subsequences, including λ and x itself.

Now we will use the concept of subsequence to construct a vector-representation $\mathbf{x}$ for the sequence x. We do this by defining coordinates that correspond to all possible sequences that can be constructed from the alphabet Σ by setting those coordinates to 1 that correspond to sequences that occur as a subsequence in $x = abac$

u:	λ	a	b	c	aa	ab	...	cc	aaa	...	aba	...
$r(u)$:	0	1	2	3	4	5	...	12	13	...	16	...
$\mathrm{x}_{r(u)}$:	1	1	1	1	1	1	...	0	0	...	1	...

Formally, from Σ, we construct the set Σ^* of all sequences that are constructible from Σ and we fix the order of the elements of Σ^*, say in lexicographical order. Then we map the ordered sequences to the nonnegative integers $\mathbb{Z}^*$, i.e. each sequence $u \in \Sigma^*$ is mapped to an integer $r(u) \in \mathbb{Z}^*$ and we use these integers to index the coordinates of the vectors. So, for each sequence x, we construct a binary vector $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \dots)$ such that

$$\mathrm{x}_{r(u)} = \begin{cases} 1 & \text{if } u \sqsubseteq x \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

This construction characterizes strings by their subsequences and the resulting vectors are also called "feature vectors", the subsequences being treated as features of the sequence. However, since the number of sequences that can be constructed from even a small alphabet is countably infinite, i.e. as big as the size of the set of nonnegative integers, actually constructing the vectors and calculating their products as required in Eq. 4.5 is not feasible. Therefore, one needs special methods, called "kernels" (see e.g. Schölkopf and Smola 2002; Shawe-Taylor and Cristianini 2004) to calculate the quantities appearing in Eq. 4.5.

Once the principle of assigning feature vectors to sequences is understood, it is easy to generalize and construct representations that are far more sophisticated than the simple binary vector as sketched above. For example, we may want to account for the length, the duration or the embedding frequency or a combination of such properties. This accomplished by a generalized representation:

$$\mathrm{x}_{r(u)} = \begin{cases} f(u,x) & \text{if } u \sqsubseteq x \\ \\ 0 & \text{otherwise} \end{cases}. \tag{4.7}$$

Provided a suitable kernel function can be found, any specification of the function f in the above representation may be used (e.g. Elzinga and Wang 2011).

Axioms of Similarity

The concept of metric distance is an old, well established concept in all sciences and there is no debate about its definition or usefulness. However, although the concept of similarity is widely used in many branches of science and engineering, especially in biological taxonomy, in chemistry and in psychology, a widely accepted definition is still lacking and there is an abundance of authors (see e.g. Batagelj and Bren 1995; Gower 1971; Gower and Legendre 1986; Holliday et al. 2002; Wang 2006) proposing quite different, application-specific quantifications of similarity.

Because of the widely different applications of similarity, it is not very sensible to propose another similarity measure for comparing sequences in the social sciences. Rather, it is much more interesting to formulate a set of intuitive axioms that quantifications of similarity should adhere to, irrespective of the application. Therefore, we will discuss a proposal recently made in Chen et al. (2009) and generalized in Elzinga et al. (2008) and in subsequent subsections discuss normalization and some relations between a similarity and a distance. Finally, we will mention some similarities that could be used in combination with some of the well-known distance metrics for sequences like OM and subsequence-based distance.

Similarity

Intuitively, two objects are similar if they share one or more features or properties and similarity seems to increase with an increasing number of such shared features. What relevant features are and whether or not all relevant features are equally important depends on the application area and the purpose of the comparison. Let X denote the set of features or properties possessed by some object x and let Y denote the set of features of object y. When we interpret similarity $s(x, y)$ as the amount of common, shared features, we are inclined to imagine

$$s(x, y) = |X \cap Y|, \tag{4.8}$$

i.e. we interpret similarity as the size $|X \cap Y|$ of the intersection $X \cap Y$ of the feature sets X and Y. With this interpretation of similarity, the next axioms directly follow from Eq. 4.8:

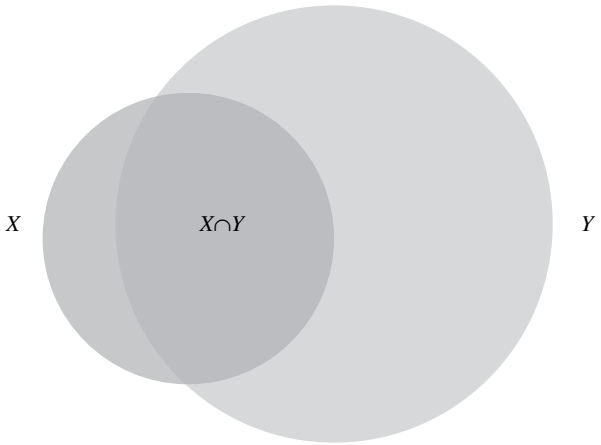

Fig. 4.3 When we assign to sequences x and y associated sets of features X and Y, $s(x, y)$ could be interpreted as the (weighted) number of common, shared features, i.e. as if $s(x, y) = |X \cap Y|$

$$S1 : s(x, y) \geq 0,$$
$$S2 : \min\{s(x, x), s(y, y)\} \geq s(x, y),$$
$$S3 : s(x, y) = s(y, x),$$

$S1$ follows because a number of common features, a count, cannot be negative. $S2$ follows because there can be no more shared features than possessed by either of the objects and $S3$ follows from the symmetry of intersection $X \cap Y = Y \cap X$. These principles are illustrated in Fig. 4.3. The reader notes that the first three axioms are highly similar to the distance axioms $D1 - D3$. Only, $D1$ says that the distance $d(x, x)$ is minimal while $S2$ states that the similarity $s(x, x)$ is maximal. So, it seems that distance and similarity are opposite counterparts. In one of the next subsections, we scrutinize the relation between distance and similarity.

What is lacking from the similarity axioms is an axiom that bounds the similarity function like the triangle inequality $D4$ bounds the distances. However, when we look at Fig. 4.4, the axiom

$$S4 : s(x, y) + s(y, z) \leq s(x, z) + s(y, y)$$

seems inevitable and we will call this inequality the "covering inequality" because when $s(x, z) = 0$, $s(x, y) + s(y, z)$ cannot exceed $s(y, y)$. Again, the reader notes that the direction of the covering inequality is opposite to that of the triangle inequality. Finally, if we operationally define object equality $x = y$ if and only if $|X| = |Y|$, i.e. when we define object equality through equality of feature sets, we must have that

$$S5 : s(x, x) = s(y, y) = s(x, y) \text{ if and only if } x = y$$

and this axiom connects similarity to object-equivalence just like $D1$ connects distance to object-equivalence. The reader notes that, just like the distance axioms do not prescribe how to measure, how to establish distances in space, the similarity

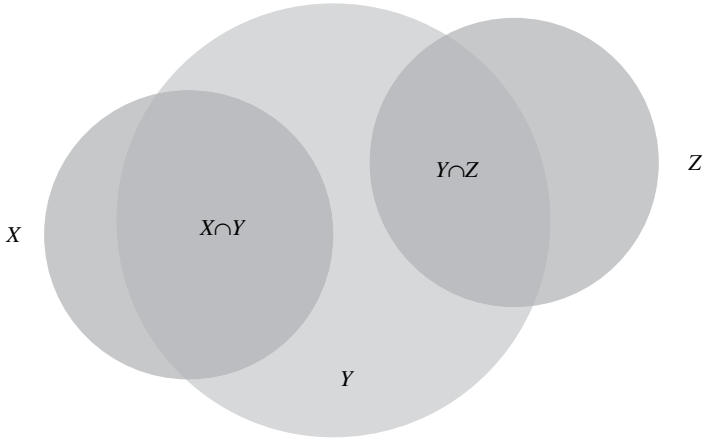

Fig. 4.4 Similarity interpreted as the size of common subsets of features and self-similarity $s(x,x) = |X|$. In the Venn-diagrams below, we have that $s(x,y),\ s(y,z) > 0$ but $s(x,z) = 0$ violates $s(x,y) + s(y,z) \leq s(x,z)$ but the general rule $s(x,y) + s(y,z) \leq s(x,z) + s(y,y)$ will always hold

axioms do not prescribe how to establish similarity. Also, it is important to note that $s(x,y) = X \cap Y$ is an interpretation of s but there is no need to actually construct a measure s through comparisons of feature sets. Actual measuring systems should only adhere to the axioms in order that they yield metric distance or similarity.

On the Wrong Track

In several papers (e.g. in Bras et al. (2010) and in Elzinga and Liefbroer (2007)) we used a subsequence-based vector-representation of life course sequences and proposed to use

$$s(x,y) = \frac{\mathbf{x}'\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{x} \cdot \mathbf{y}'\mathbf{y}}} \tag{4.9}$$

as a similarity measure. This $s(x,y)$ satisfies the similarity axioms S1–S3 and S5 and it has the nice, additional property that its numerical value is easy to interpret since $0 \leq s(x,y) \leq 1$ (actually, $s(x,y)$ evaluates the cosine of the angle between the representing vectors $\mathbf{x}$ and $\mathbf{y}$). Unfortunately, as a general measure of similarity between vectors, $s(x,y)$ does not satisfy the covering inequality S4. This is easily demonstrated with a toy example[2]: suppose that we have vectors $\mathbf{x} = (0,1)$, $\mathbf{y} = (1,0)$ and $\mathbf{z} = (1,1)$. Then the matrix of similarities according to Eq. 4.9 is given by

[2] This example was suggested to me by Matthias Studer through personal communication.

	x	y	z
x	1		
y	0	1	.
z	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	1

According to the covering inequality, we should have that

$$s(x, y) + s(z, z) \geq s(x, z) + s(z, y)$$

but we observe that

$$0 + 1 < \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2} \approx 1.41,$$

a clear violation of the covering inequality. Of course, this is only a toy-example. But the example demonstrates that we cannot be certain that $s(x, y)$ calculated as defined by Eq. 4.9 for real-life vector-products satisfies S4.

Back on Track Again

We began our thinking about similarity from the intuition that similarity should be proportional to the size of the set of common features. This invites to define

$$s(x, y) = |X \cap Y| \tag{4.10}$$

wherein X and Y denote the sets of features of the objects x and y. Indeed, it is true that this $s(x, y)$ satisfies the similarity axioms S1-S5 (for a proof, see Chen et al. (2009) or Elzinga et al. (2008)). Let us now look back at the vector representation of Eq. 4.6, which we repeat here for convenience:

$$\mathrm{x}_{r(u)} = \begin{cases} 1 & \text{if } u \sqsubseteq x \\ 0 & \text{otherwise} \end{cases}. \tag{4.11}$$

When we say that the subsequences of x constitute the feature set X of x, then we have that

$$\mathbf{x}'\mathbf{x} = \sum_i \mathrm{x}_i^2 = |X|, \tag{4.12}$$

$$\mathbf{x}'\mathbf{y} = \sum_i \mathrm{x}_i \mathrm{y}_i = |X \cap Y| \tag{4.13}$$

and hence that the vector-product itself satisfies the similarity axioms S1-S5. So, we could set $s(x, y) = \mathbf{x}'\mathbf{y}$ in order to obtain a proper similarity for a vector-space. However, we then have two practical problems. The first is that the numerical value of this $s(x, y)$ is hard to interpret: if it would equal 712340762134, would that mean that the objects x and y are very much alike? We simply wouldn't know unless we knew the value of s of all pairs of objects. The second problem is that we may encounter very many pairs of objects x and y for which $s(x, x) \neq s(y, y)$, simply because $|X| \neq |Y|$. This would be counterintuitive since we are inclined to think that all objects are equally similar to themselves.

To remedy these problems of interpretation, we would like to see that s is tightly bounded, preferably by 0 and 1, and that $s(x, x) = s(y, y)$ for all pairs of objects. But this was exactly the purpose of the definition of $s(x, y)$ in Eq. 4.9, the "cosine-similarity". Apparently, not all normalizations of a proper similarity generate a proper normalized similarity. Therefore, in the next sections, we will turn our attention to properly transforming and normalizing distances and similarities.

Units and Transformations

Distances are not dimensionless numbers. Instead, distances are expressed in terms of some standard unit of length. When we consider distances in our physical environment, we always mention the unit of length that the numbers refer to: kilometers, nautical miles, lightyears or Ångströms and what not.

When we calculate distance between sequences, we also have a unit of distance although we tend not to mention it. However, even the simple Hamming-distance (Hamming 2010) has a unit: it counts the number of positions in which two sequences have unequal states. Hence, Hamming's unit of distance is "position". Similarly, the unit of OM-distances, when indel cost is set to 1 and substitution cost is set to 2, refers to the number of "edits". When the OM-distance $d_{OM}(x, y) = 10$, it means that the minimum number of edits to change x into y equals 10. So, comparison of distances is straightforward, even if the pertaining pairs of sequences are defined over different state alphabets.

However, there are at least two kinds of problems that warrant choosing a different scale for our distances. First, when the distances are very big numbers, it may be advisable to chose a bigger unit of distance. This will facilitate the interpretation of e.g. averages and standard deviations within clusters or the averages and standard deviations of distances to centroids or medoids of clusters or to particular "prototypes". Second, certain sequences may be extremely remote from other sequences in the data and such extreme data may heavily influence the subsequent clustering or discrepancy analysis of the sequences. In such cases, we would be well-advised to apply a compressive but order-preserving transformation, e.g. a logarithmic transformation, to the distances as calculated before further analysis. Hence, we have to talk about admissible transformations $f(\,.\,)$ of the form $d'(x, y) = f(d(x, y))$. Of course, admissible transformations are those that, if d is a metric, ensure that $d' = f(d)$ is a distance metric too, i.e. ensure that d' too satisfies the four metric axioms.

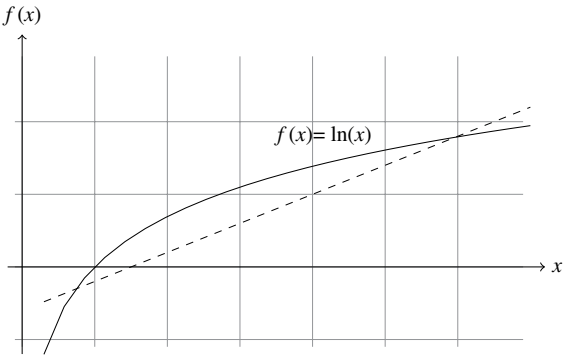

Fig. 4.5 Plot of the concave function $f(x) = \ln(x)$. If there are two intersections of $f(x)$ with the same straight line, then between the intersections, the graph of $f(x)$ is above the straight line

All admissible transformations f satisfy a few simple properties:

$$A1 : f(0) = 0,$$
$$A2 : a < b \text{ if and only if } f(a) < f(b),$$
$$A3 : f(a+b) \leq f(a) + f(b),$$

for all real numbers a and b. Clearly, $A1$ ensures that the interpretation of 0-distance (object-equivalence) is retained, $A2$ ensures that the order of distances is retained (monotonicity) and properties $A2$ and $A3$ (sub-additivity) together ensure that the triangle inequality is retained after transformation of the distances. If we require that f is continuous, such an f must be *concave*. A graphical account of what "concave" means is given in Fig. 4.5 where we use $f(x) = \ln(x)$ as an example. To see if a function is concave, draw its graph and a straight line intersecting it at two locations. If the graph is not below any such straight line, then we say that the function is concave. The reverse case, when the graph is nowhere above the straight line between the points of intersection, we say that the function is *convex*. A limiting case is the linear function: it is both concave and convex. Concave functions all have the property that the growth of the function values decreases everywhere: in Fig. 4.5, the increase of $f(x)$ always diminishes with increasing x.

So, examples of admissible transformations are (see also Batagelj and Bren 1995)

$$\begin{array}{lll} f(x) = ax & \text{with} & a > 0, \\ f(x) = \log_c(x+1) & & \\ f(x) = x^p & \text{with} & 0 < p \leq 1, \\ f(x) = a^{-x} - 1 & \text{with} & a > 0 \end{array}$$

Why do we only allow for concave transformations? Why do we not accept all order-preserving transformations? The reason is that we want the transformed distances to satisfy the triangular inequality too:

$$f(d(u,w)) \leq f(d(u,v)) + f(d(v,w)).$$

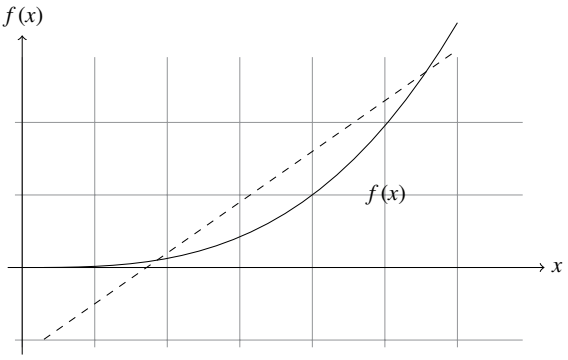

Fig. 4.6 Plot of a convex function $f(x)$. If there are two intersections of $f(x)$ with the same straight line, then between the intersections, the graph of $f(x)$ is below the straight line

This can only be attained when f is non-decreasing in such a way that the left side of the inequality does not grow faster than the sum in the right side. For example, suppose that

$$d(u,w) = 10,\ d(u,v) = 9\ and\ d(v,w) = 2$$

and that we would set $f(x) = x^2$. This would yield

$$f(d(u,w)) = 100 > f(d(u,v)) + f(d(v,w))$$

and thus we would violate the triangular inequality. The covering inequality $S4$ is the opposite of the triangular inequality, hence admissible transformations of similarities should have the opposite quality of concavity: convexity. A graphical definition of that property is provided in Fig. 4.6. Precisely: if f is a convex function such that $f(0) \geq 0$ and $f(x) < f(y)$ whenever $x < y$, then f is an admissible similarity transformation, i.e. if $s(x,y)$ is a similarity, then $s'(x,y) = f(s(x,y))$ is a similarity[3] too. Examples of admissible similarity transformations are

$$f(x) = \alpha x^p + \beta \text{ with } \alpha > 0, \beta \geq 0 \text{ and } p \geq 1, \tag{4.14}$$

$$f(x) = e^x. \tag{4.15}$$

Normalization

A weight-difference of 10 kg between male adults of roughly the same age and length may not be very significant but that very same weight difference is a matter of life and death when it pertains to two 2-year old children. Similarly, a difference

[3] Chen et al. (2009) use the expression "similarity metric" for any s that satisfies the axioms S1-S5. This is well-defendable since a similarity s "metricizes" the sequence space, just like a distance. However, we prefer to call such an s a "similarity" since the noun "metric" has been associated with "distance" for over a century now.

of 2 years of unemployment between careers of over 30 years on the labor market may be insignificant, but the same difference between the labor market careers of two 18-year old youngsters could have a dramatic differential effect on their future careers.

So, in many applications of measurement, differences or distances between pairs of objects are weighted according to the properties of the separate objects. As soon as we start considering differences, distances or proximities relative to the properties or features of the objects involved, we will almost automatically do two different things: we will tend to use relative, *dimensionless* or unit-free measures and we will create a *bounded* scale. Bounded by the maximum and/or the minimum of the difference, distance or proximity that could have been obtained, given the properties of the pertaining objects. For example, the weight difference between two male adults probably cannot exceed an upper bound of 250 kg. So, expressing the actual, observed difference relative to this maximum, will convey useful information about an observed weight difference.

The above intuitions about dimensionlessness and boundedness are captured by the notion of normalization. We will say that a measure M is "normalized" precisely when it satisfies the next two properties:

1. M is tightly bounded, i.e. $a \leq M \leq b$ for some numbers $a < b$,
2. M is dimensionless or, equivalently, unit-free.

Some authors require only boundedness and do not demand the dimensionlessness. When a measure is bounded, we know that its maximum and minimum values are fixed, independent of the (pairs of) objects it is applied to. We say that the bounds are "tight" when the maximum and minimum values, the boundaries, will be actually attained if M is applied to some suitable, real (pair of) object(s). A good example of a tightly bounded measure is Pearson's correlation coefficient r: we know that $-1 \leq r \leq 1$ and that either of these boundaries will indeed be attained when the one variable is a linear transform of the other variable. The value 1.6 is a boundary of r too since $r < 1.6$ but 1.6 is not a tight boundary: $1 < 1.6$ is the smallest, the tightest upper boundary.

When a measure is "dimensionless", we know that the numerical value of the measure does not refer to a unit or, equivalently, is not affected by admissible transformations of the scale it derives from. Again, Pearson's r is a good example since linear transformations of either variable will not affect the numerical value of the measure of association $M = r$.

These two properties, boundedness and dimensionlessness, are valuable properties: boundedness implies that we can interpret the actual value of the measure with respect to its boundaries ($r = 0.87$ is a "high" correlation since close to the upper bound of 1) and we can compare different instances, different values of the measure, independently of sample space or the scales of the pertaining variables ($r = 0.87$ denotes a stronger degree of linear association than $r = 0.43$, regardless of the scales involved).

Therefore, normalization of a measure will greatly enhance its applicability and practical usefulness and thus, it makes sense to report normalized versions of measures, in particular of measures of distance and similarity. However, we should

be aware of the fact that a normalizing transform may not be order-preserving: a weight-difference of 10 kg for 2-year old kids is very serious, much more serious than a 12 kg weight-difference between two adults and this will be expressed in the normalized versions of the weight differences. So, a normalized distance or similarity does *not* preserve the order of the original distances since it weights relative to the properties of the pertaining objects. When these properties differ, the same weight will be normalized to different values, depending on the pertaining pairs of objects.

As distance and similarity are nonnegative, it is practical to normalize such that our measures will be tightly bounded by the closed interval [0,1]. A tight upper bound of 1 will invite to interpret the actual values as fractions of the maximum attainable upper bound. Normalization of distance or similarity should not lead to a loss of one of the metric properties. So, we demand that a normalized distance D and a normalized similarity S satisfy the axioms as stated below:

Distance	Similarity
D′1 $D(x, x) = 0$	S′1 $S(x, x) = 1$
D′2 $0 \le D(x, y) < 1$	S′2 $1 > S(x, y) \ge 0$
D′3 $D(x, y) = D(y, x)$	S′3 $S(x, y) = S(y, x)$
D′4 $D(x, z) \le D(x, y) + D(y, z)$	S′4 $S(x, z) + 1 \ge S(x, y) + S(y, z)$

Clearly, the two axiom systems are complementary so we expect that normalized distances can be obtained from normalized similarities and vice versa. Indeed, we will make some remarks on such conversions later. First, we will turn to the most important problem, the actual normalization of distance and similarity, such that the results adhere to the axioms D'1-D'4 or S'1-S'4 as stated above. In passing, we will specifically deal with OM-based and subsequence-based metrics.

Normalized Distance

Here, we discuss two most simple versions of distance-normalizations as proposed by Chen et al. (2009) and apply them to OM- and subsequence-based metrics.

The first method relies on a simple "bounding" transform:

$$f_b(x) = \frac{x}{x+b} \text{ for } x \ge 0 \text{ and } b > 0 \tag{4.16}$$

Clearly, when x=0, $f(x) = 0$ and for increasing values of x, $f(x)$ will tend to 1 since whatever the value of b, it will become almost irrelevant for a big enough x. In Fig. 4.7, we show curves for four different values of b. A transformation like f_b has two shortcomings: when applied to a distance, the resulting transformed distance does not satisfy the triangle inequality and it does not yield a dimensionless quantity since it does not relate to the properties of the individual objects. To remedy these

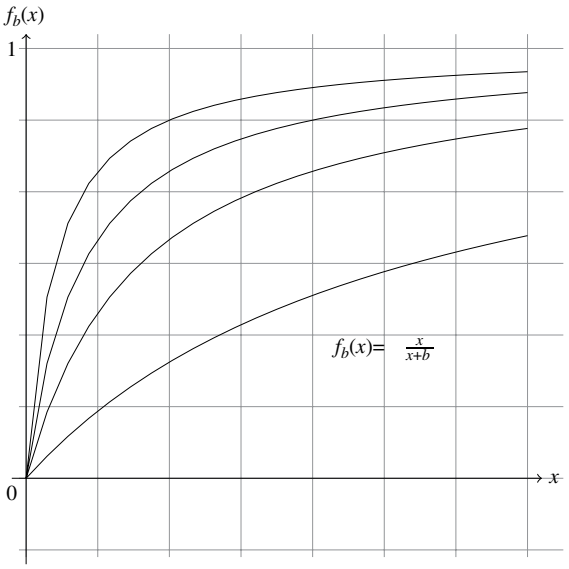

Fig. 4.7 Plots of normalizing transform for $b = \{0.4, 0.8, 1.6, 5.4\}$

shortcomings, we replace the bounding constant b by a quantity that pertains to the individual objects involved and that ensures that the resulting quantity does satisfy the triangle inequality:

$$0 \leq D_r(u,v) = \frac{d(u,v)}{\{d(u,v) + \underbrace{d(u,r) + d(r,v)}_{b}\}/2} \leq 1 \tag{4.17}$$

is a normalized distance for any reference object r. Often, it is most convenient to set $r = \lambda$, i.e. to take the empty sequence as the reference object. That 1 is the smallest upper boundary of D_r derives from the fact that d is a distance and thus satisfies the triangle inequality: since $d(u,v) \leq d(u,r) + d(r,v)$, we must have that $(d(u,r) + d(r,v) + d(u,v))/2 \geq d(u,v)$ and thus that $D_r(u,v) \leq 1$. Whatever the reference object r, $D_r(u,v)$ will depend on it. But $D_r(u,v)$ will not only depend on r and the comparison of u and v, but also on the comparison of u with r and the comparison of v with r. This implies that D_r is *not* order-preserving. For example, if $d(u,v) = d(u',v')$ but u' and v' are much more remote from r than u and v, $D_r(u,v) > D_r(u',v')$ (The reader is invited to graphically display this example). Hence, "raw" distances $d(u,v)$ are weighed by the lengths of the sequences involved. At first sight this may seem to be an unfortunate state of affairs. However, let us apply Eq. 4.17 to standard-cost OM-distance d_{OM} and set $r = \lambda$, i.e. take the empty sequence as the reference object. Then the above normalisation comes down to

$$D_\lambda(u,v) = \frac{d_{OM}(u,v)}{(d_{OM}(u,v) + |u| + |v|)/2} \tag{4.18}$$

wherein $|u|$ denotes the length of the sequence u. Now suppose that $d_{OM}(u,v) = 5$ and $|u| = 7 = |v|$. When we have to use at least 5 edits to turn a short u into a short v, $d(u,v)$ is a big distance. But if $|u'| = 50 = |v'|$, $d(u',v') = 5$ is only a small distance and we will see that $D_\lambda(u,v) > D_\lambda(u',v')$. So indeed, "raw" distances are weighed by the lengths of the sequences involved and this attractive property results in a normalized distance that is *not* order-preserving. The same effect will occur when we apply Eq. 4.17 to subsequence-based distances d_V:

$$D_\lambda(u,v) = \frac{d_V(u,v)}{(d_V(u,v) + \sqrt{\mathbf{u'u}} + \sqrt{\mathbf{v'v}})/2}. \tag{4.19}$$

In both cases, the original distances of the sequences involved are weighted according to their lengths—in the OM-metric, the length of u relative to λ equals $|u|$ and in the subsequence-metric it equals $\sqrt{\mathbf{u'u}}$. In actual application, it is relevant to stress that Eq. 4.18 is correct only when the standard cost matrix is applied. If not, we have to calculate $d_{OM}(u,\lambda)$ as the sum $\delta(u)$ of the deletion costs of all states of u and thus our general normalizer for OM-distances becomes

$$D_\lambda(u,v) = \frac{d_{OM}(u,v)}{(d_{OM}(u,v) + \delta(u) + \delta(v))/2}, \tag{4.20}$$

a normalization already proposed by Yujian and Bo (2007). A second way to normalize distances is through

$$D_r(u,v) = \frac{d(u,v) - \min\{d(u,r), d(r,v)\} - \max\{d(u,r), d(r,v)\}}{2 \cdot \max\{d(u,r), d(r,v)\}} \tag{4.21}$$

Again, this D_r is not order-preserving, due to the same effects as explained above and it can be adapted in an analogues way to OM- or vector-based distances. The details of applying this normalization to OM or vector-based distance are left to the reader. A simple normalization that does not depend upon a reference object r but satisfies the axioms D'1-D'4, is attained by using the exponential transform as in

$$D(u,v) = 1 - e^{-d(u,v)}. \tag{4.22}$$

Consequently, the latter normalization will *not* weigh distance according to lengths of the sequences involved, it is order-preserving and $D(u,v) = D(u',v')$ whenever $d(u,v) = d(u',v')$. It is not a proper normalization since the result is not unit-free. I find it difficult to imagine a sensible application of it. For OM, normalizations have been proposed, e.g. in Gabadinho et al. (2011), that are quite similar to the normalizers that we have seen so far:

$$D(u,v) = \frac{d_{OM}(u,v)}{|u| + |v|} \text{ and } D(u,v) = \frac{d_{OM}(u,v)}{\max\{|u|, |v|\}} \tag{4.23}$$

but these normalising transformations do not yield proper distances in the sense that they not adhere to the triangular inequality. Therefore, the use of such normalizers should be avoided.

Similarities and Their Normalization

So far, we discussed general properties of similarities but we did not discuss how to construct them in actual practice. This is especially relevant for OM-based analysis, since the OM-algorithm, provided with a metric cost matrix, generates distances. So, these distances must serve as the basis for the construction of edit-based similarity. Therefore it is relevant to mention two different ways to construct a similarity, given some distance metric d. Both methods require a reference object r; once this has been set, we have that

$$s_1(u,v) = d(u,r) + d(v,r) - d(u,v) \tag{4.24}$$

and

$$s_2(u,v) = \min\{d(u,r), d(v,r)\} - d(u,v). \tag{4.25}$$

and s_1 and s_2 both satisfy the similarity axioms S1–S5. Both constructions lead to objects that are remote from r being more similar than objects that are close to r. We mention some details pertaining to s_1. First, we deal with the general OM-variant of it, setting $r = \lambda$. This yields

$$s_1(u,v) = \delta(u) + \delta(v) - d_{OM}(u,v) \tag{4.26}$$

wherein $\delta(u)$ again denotes the sum of all deletion costs of the characters of u. Interestingly, Yujian and Bo (2007) proposed a special case of the above similarity and then proved that it satisfies the covering inequality S4 but they did not recognize that inequality as a general property of similarity measures. For vector-based distances and $r = \lambda$, Eq. 4.24 yields

$$s_1(u,v) = \sqrt{\mathbf{u}'\mathbf{u}} + \sqrt{\mathbf{v}'\mathbf{v}} - d(u,v). \tag{4.27}$$

Similar details for s_2 are left to the reader. There is a well-known similarity coefficient for sets, first proposed by Rogers and Tanimoto (1960): for sets X and Y, the quantity

$$0 \leq T(X,Y) = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \leq 1. \tag{4.28}$$

is a normalized similarity measure. The coefficient is widely known as the Tanimoto-coefficient (see e.g. Duda et al. 2001) and several authors have come up with

generalizations and variants (see e.g. Tversky 1977). Lipkus (1999) proved that the Tanimoto-coefficient satisfies the covering inequality S4 and used this to prove that $D(X, Y) = 1 - T(X, Y)$ is a distance metric over sets. A more general formulation of the Tanimoto-coefficient yields a normalizer for any similarity and thus for the coefficients specified in Eqs. 4.24 and 4.25:

$$S_1(u, v) = \frac{s(u, v)}{s(u, u) + s(v, v) - s(u, v)} \tag{4.29}$$

An alternative normalizer is

$$S_2(u, v) = \frac{s(u, v)}{\max\{s(u, u), s(v, v)\}} \tag{4.30}$$

More and more general normalizers are discussed in Chen et al. (2009). Detailing S_1 for the case of a vector-representation yields the normalizer

$$S_1(u, v) = \frac{\mathbf{u}'\mathbf{v}}{\mathbf{u}'\mathbf{u} + \mathbf{v}'\mathbf{v} - \mathbf{u}'\mathbf{v}}. \tag{4.31}$$

Finally, it is always possible to generate a normalized similarity S from a normalized distance D, since whichever of S or D is given, the other can be easily obtained through the equation $S(u, v) = 1 - D(u, v)$.

Similarity or Distance: does it Matter?

A final issue to be dealt with relates to the generally held belief that distance and similarity are interchangeable. For example, Martin et al. (2008) write:

> In the context of the efforts to evaluate similarity and change in life course transitions, optimal matching analysis (OMA) has been recommended to complement classical statistical methods to make use of the holistic information encoded in biographical status sequences.

Similar neglecting of the difference between distance and similarity is abundant (e.g. in Gauthier et al. 2010) and not only in the social sciences. The belief seems to be that distance and similarity are opposite or inverse in the sense that similar sequences should be close, that dissimilar sequences should be remote, that remote sequences should be dissimilar, etc. However, distances and similarities are just vehicles that are used to create partitions of big sets of sequences and it is not at all clear that partitioning on the basis of distance will lead to the same partitioning as partitioning on the basis of similarities.

Interestingly, Emms and Franco-Penya (2013) investigated precisely this problem, confined to edit-based distances and similarities. Remarkably, one of their conclusions is that partitions based on hierarchical clustering of distances can always

Table 4.1 *d* denotes an arbitrary distance metric and *s* an arbitrary similarity measure and *S* and *D* denote their normalized versions. The table shows how to transform a row quantity to a column quantity, either by providing a simple formula, or by referring to equation numbers from this chapter

	d	*s*	*D*	*S*
d	concave *f*	4.24, 4.25	4.17, 4.21, 4.22	e^{-d}
s		convex *f*		4.29, 4.30
D	concave *f*			1-*D*
S		convex *f*	1-*S*	

be replicated by similarity-based hierarchical clustering but not vice versa: some similarity-based clusterings will not be replicable through distance-based hierarchical clustering. This suggests that similarity is a more general, more encompassing concept than distance. Indeed, this is reflected in the axiom systems: the negation of the axioms of a distance yields a similarity but the negation of the axioms of a similarity does not necessarily yields a distance.

Summary

In this chapter, I focussed on the concepts of distance and similarity and their intricate relations. I tried to explain the importance of the axiomatic foundation of the concepts and the importance of regularity-axioms like the triangle inequality and the covering inequality. I also tried to explain the principles of admissible transformations and, more importantly, the principle and consequences of normalization. In passing, I provided for some ready-to-use implementations for OM- and vector-based distances. To help the reader find her way through the forest of intricate relations and formulae, I constructed Table 4.1: it shows how to transform a "row"-measure into a "column"-measure. Some of the cells of Table 4.1 are empty; not because such transformations are impossible but because I could not think of a sensible application. The reader should be aware that *none* of the transformations I described is unique in the sense that they would be the only solutions to the pertaining transformation-problem; we only know that these solutions respect the axioms of distance or similarity and there might be a wealth of other solutions.

In the last section, I made a few remarks on the fact that similarity is a more encompassing, more general concept than distance. So we have to be precise in the questions we ask and the concepts we use. Once a structuring concept—distance or similarity—is chosen, the next question is to either embed in an OM-space or a vector-space. Finally, the issue to be dealt with is that of a cost-matrix or, in case of a vector-representation, the features to incorporate and how to weigh and compare them.

References

Batagelj, V., & Bren, M. (1995). Comparing resemblance measures. *Journal of Classification, 12*, 73–90.

Berghammer, C. (2010). Family life trajectories and religiosity in Austria. *European Sociological Review, 26*, 1–18.

Bonetti, M., Piccarreta, R., & Salford, G. (2013). Parametric and nonparametric analysis of life courses: An application to family formation patterns. *Demography, 50*, 881–902.

Bras, H., Liefbroer, A. C., & Elzinga, C. H. (2010). Standardization of pathways to adulthood? An analysis of Dutch cohorts born between 1850 and 1900. *Demography, 47*(4), 1013–1034.

Chen, S., Ma, B., & Zhang, K. (2009). On the similarity metric and the distance metric. *Theoretical Computer Science, 410*(24–25), 2365–2376.

Crochemore, M., Hancart, C., & Lecroq, T. (2007). *Algorithms on strings*. Cambridge: Cambridge University Press.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: Wiley.

Elzinga, C. H. (2003). Sequence similarity—A non-aligning technique. *Sociological Methods and Research, 31*(4), 3–29.

Elzinga, C. H. (2005). Combinatorial representation of token sequences. *Journal of Classification, 22*(1), 87–118.

Elzinga, C. H., & Liefbroer, A. C. (2007). De-standardization and differentiation of family life trajectories. *European Journal of Population, 23*(3–4), 225–250.

Elzinga, C. H., & Wang, H. (2013). Versatile string kernels. *Theoretical Computer Science,495*, 50–65.

Elzinga, C. H., Rahmann, S., & Wang, H. (2008). Algorithms for subsequence combinatorics. *Theoretical Computer Science, 409*(3), 394–404.

Elzinga, C. H., Wang, H., Lin, Z., & Kumar, Y. (2011). Concordance and concensus. *Information Sciences, 181*, 2529–2549.

Emms, M., & Franco-Penya, H.-H. (2013). On the expressivity of alignment-based distanceand similarity measures on sequences and trees in inducing orderings. In P. Latorre Carmona, J.J. Sánchez, & A.L. Fred (Eds.), *Mathematical methodologies in pattern recognition and machine learning. ICPRAM 2012 international conference on pattern recognition applications and methods, vol. 30 of Springer proceedings in mathematics & statistics*, (pp. 1–18). New York: Springer.

Fasang, A. E. (2010). Retirement: Institutional pathways and individual trajectories in Britain and Germany. *Sociological Research Online, 15*(2), 1.

Fréchet, M. R. (1906). Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo, 22*(1), 1–72.

Gabadinho, A., Ritschard, G., Müller, N., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR.*Journal of Statistical Software, 40*(4), 1–37.

Gauthier, J.-A., Widmer, E., Bucher, P., & Notredame, C. (2010). Multichannel sequence analysis applied to social science data. *Sociological Methodology, 40*(1), 1–38.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27*, 857–871.

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification, 3*, 5–48.

Halpin, B. (2010). Optimal matching analysis and life-course data: The importance of duration. *Sociological Methods & Research, 38*(3), 365–388.

Hamming, R. W. (1950). Error-detecting and error-correcting codes. *Bell System Technical Journal, 26*(2), 147–160.

Han, S.-K., & Moen, P. (1999). Clocking out: Temporal patterning of retirement. *American Journal of Sociology, 105*(1), 191–236.

Holliday, J. D., Hu, C.-Y., & Willett, P. (2002). Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D-fragment bit-strings. *Combinatorial Chemistry and High-Throughput Screening, 5*, 155–166.

Hollister, M. N. (2009). Is optimal matching sub-optimal? *Sociological Methods & Research, 38*, 235–264.

Lipkus, A. H. (1999). A proof of the triangle inequality for the Tanimoto distance. *Journal of Mathematical Chemistry, 26*, 263–265.

Manzoni, A., Vermunt, J. K., Luijkx, R., & Muffels, R. (2010). Memory bias in retrospective collected employment careers: A model-based approach to correct for measurement error. *Sociological Methodology, 40*(1), 39–73.

Martin, P., Schoon, I., & Ross, A. (2008). Beyond transitions: Applying optimal matching analysis to life course research. *International Journal of Social Research Methodology, 11*(3), 179–199.

Massoni, S., Olteanu, M., & Rousset, P. (2009). Career-path analysis using optimal matching and self-organizing maps. In J.C. Príncipe & R. Miikkulainen (Eds.), *Advances in self-organizing maps. Lecture notes in computer science 5629*. (pp. 154–612). New York: Springer.

Rogers, D. H., & Tanimoto, T. T. (1960). A computer program for classifying plants. *Science, 132*, 1115–1118.

Rohwer, G., & Pötter, U. (1999). *TDA User's Manual*. Bochum: Ruhr-Universität.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels. Support vector machines, regularization optimization, and beyond*. Cambridge: MIT Press.

Shawe-Taylor, J., & Christianini, N. (2004). *Kernel methods for pattern recognition*. Cambridge: Cambridge University Press.

Studer, M., Ritschard, G., Gabadinho, A., & Muller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research, 40*(3), 471–510.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–352.

Wang, H. (2006). Nearest neighbors by neighborhood counting. *IEEE Transactions on Pattern Learning and Machine Intelligence, 28*(6), 1–12.

Yujian, L., & Bo, L. (2007). A normalized Levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(6), 1091–1095.